Stato del Computing in Belle II

Dr. Silvio Pardi

Meeting dei Siti Italiani

06 Dicembre 2024

KEKCC Renewal

- Unexpected long downtime
  - Metadata service took ~2.5 weeks to be migrated (Down Mon 8/26 - Thu9/12)
  - No same issue in future
  - We plan to move away from this "old" technology (AMGA) to Rucio
- Less resources available after the renewal
- Belle II servers in the new KEKCC delivered late
- DIRAC Migrated to the new server (October 2024)

Migration to EL9

Sites are migrating to EL9 their infrastructure. Activity monitored as the following pages

https://people.na.infn.it/~spardi/CE-status.html

- Jobs using basf2 release-6 (or the old light-22xx) cannot run on EL9
- Not only the grid resources, but the interactive work servers, too
- KEKCC and DESY NAF are now EL9
- Libraries compiled on EL9 cannot be used EL7/EL8
- release-6 cannot run on EL9, but it works on EL8 but may not work for user jobs with private libraries built on CentOS7.

Migration to EL9

Basf2 software of belle II has a set of dependencies we are discussing which ones will be distributed via CVMFS.

For EL8/EL9 Worker Node we have the following

wn-5.1.0-1 from WLCG repository Required

HEP_OSlibs-9.2.2-3 from WLG repository Recommended

In addition site that migrate to EL9 generally migrate to the new version of Condor, this require a re-certification of the site, the change fo CE on DIRAC, dismiss the old CE and troubleshooting the various issues.

Apptainer implementation in production

Apptainer containers are used only when a resource tag is appended to the JDL

1. At basf2helper.py, we check the OS version in the worker node
2. If the major OS version differs with the EL version in the resource tag, the `apptainer` command is prepended to the basf2 execution command
3. The image to launch the container is obtained from DIRAC CS, under Operations/Gbasf2/Containers/Images, based on the resource tag

So basically, to enable Apptainer in production, we need to add resource tags to the CEs **different than the associated with the OS** in the worker nodes (for example, `EL7` in sites with EL9).

HTCondor-CE Authentication

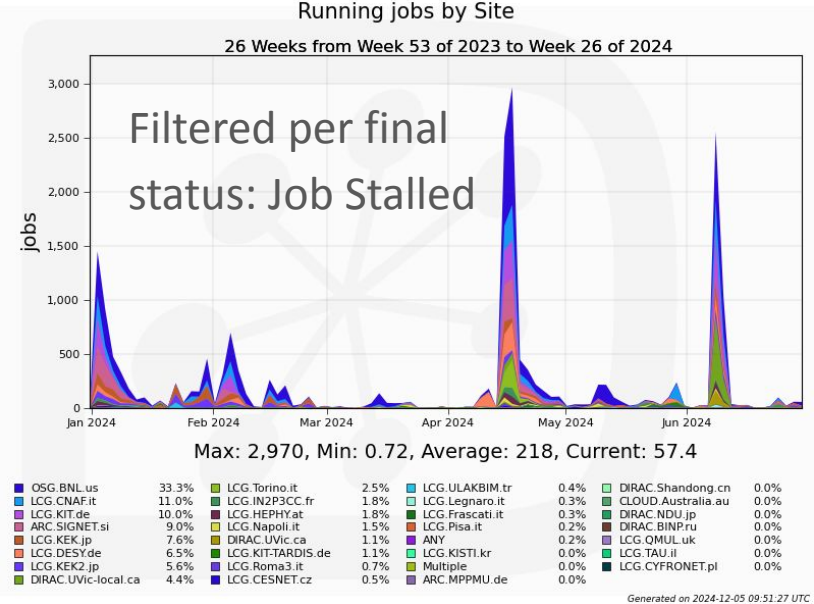
GSI Authentication is not supported anymore from the newest releases of HTCondor-CE

However Belle II is not ready for the migration to Token Based Authentication.

Currently sites which are upgrading their HTCondor should implement SSL authentication mapping the following certificates used by the pilot factories.

"/C=JP/O=KEK/OU=CRC/CN=Robot: BelleDIRAC Pilot1 - UEDA Ikuo"
"/C=JP/O=KEK/OU=CRC/CN=Robot: BelleDIRAC Pilot - UEDA Ikuo"
"/C=JP/O=KEK/OU=CRC/CN=Robot: BelleDIRAC Pilot 2 - UEDA Ikuo"

Memory Consumption by jobs



We are seeing that many production and analysis are suffering from stalled jobs.

From the current analysis it seems that this is due by the higher memory consumption (more than >2GB)

https://htcondor.readthedocs.io/en/23.x/version-history/upgrading-from-10-0-to-23-0-versions.html

I think this comment is relevant to this issue.

- In an HTCondor Execution Point started by root on Linux, the default for cgroups memory has changed to be enforcing. This means that jobs that use more then their provisioned memory will be put on hold with an appropriate hold message. The previous default can be restored by setting CGROUP_MEMORY_LIMIT_POLICY = none on the Execution points. (HTCONDOR-1974)

Can site provides more RAM per Core? They need to configure it locally?

DIRAC RequestMemory from HTCondor-CE to Condor

On LCG.Napoli.it and OSG.BNL.us sites we seen that the RequestMemory parameter from DIRAC have been sent to HTCondor-CE but not transmitted to condor cluster.

It seems due by the default configuration of HTCondor-CE

```
[root@htc-belle-ce02 config.d]# condor_ce_q 9934.0 -l |grep -i mem
MemoryProvisioned = 2048
RequestMemory = 3000
Requirements = (TARGET.Arch == "X86_64") && (TARGET.OpSys == "LINUX") && (TARGET.Disk >= RequestDisk) &&
```

Looking on the relative job on condor

```
[root@htc-belle-ce02 config.d]# condor_ce_q 9934.0 -l |grep -i RoutedTo
RoutedToJobId = "69397.0"
```

we can then query condor and we discover that RequestMemory = 2000

```
[root@htc-belle-ce02 config.d]# condor_q "69397.0" -l |grep -i mem
MemoryProvisioned = 2048
RequestMemory = 2000
```

```
[[root@htc-belle-ce02 config.d]# condor_ce_config_val JOB_ROUTER_TRANSFORM_Memory
EVALSET RequestMemory maxMemory ?: (2000)
SET JobMemory RequestMemory
if
SET JOB_GLIDEIN_Memory     "$$(TotalMemory:0)"
COPY RequestMemory OriginalMemory
COPY RequestMemory remote_OriginalMemory
SET JobMemory              JobIsRunning ? int(MATCH_EXP_JOB_GLIDEIN_Memory)*95/100 : OriginalMemory
SET RequestMemory          TARGET.TotalMemory*95/100 ?: JobMemory
endif
```

DIRAC RequestMemory from HTCondor-CE to Condor

Interaction with the INFN Condor Mailing list.
We tested this setup in Napoli which seems to solve the issue. Other site know better solution?

```
cat /etc/condor-ce/config.d/90-jobrouternew.conf

[...]

JOB_ROUTER_ROUTE_belle @=jrt
  REQUIREMENTS (x509UserProxyVoName == "belle")
  UNIVERSE VANILLA
  SET Requirements (TARGET.belle_el9_wn =?= true)
  SET Environment "BELLE2_CONDB_PROXY=http://squid-cvmfs01.na.infn.it:3128"
  SET MaxJobs 10000
  SET MaxIdleJobs 5000
  COPY RequestMemory maxMemory
@jrt

[...]
```

```
[root@htc-belle-ce02 config.d]# condor_ce_q 9950.0 -l |grep -i mem
MemoryProvisioned = 3072
RequestMemory = 3000
Requirements = (TARGET.Arch == "X86_64") && (TARGET.OpSys == "LINUX") && (TARGET.Disk >= RequestDisk) &&

[root@htc-belle-ce02 config.d]# condor_ce_q 9950.0 -l |grep -i Route
RoutedToJobId = "69497.0"

[root@htc-belle-ce02 config.d]# condor_q "69497.0" -l |grep -i mem
AutoClusterAttrs = "KFlops,MachineLastMatchTime,Offline,Rank,RemoteOwner,RequestCpus,RequestDisk,Request
JobMemory = RequestMemory
maxMemory = 3000
MemoryProvisioned = 3072
RequestMemory = 3000
```

GEANT TCS issue

The GEANT TCS relies on a back-end service provided by the commercial company SECTIGO, which handles certificate generation and renewal on behalf of the various GEANT CAs.

The company has demanded a contract renegotiation at higher price and there is an unresolved dispute over the contract's end date.

GEANT have been working hard to:

• Try to come to a new agreement with the company in question

• Look for alternative back-end providers

Due to this, it is likely that the service for user and server certificates generation and renewal will be disrupted starting from January 10, 2025.

GEANT TCS issue

We sent an email in broadcast to all belle users asking for renew the certificate before the end of the year.

Also we sent a email to the country/site rappresentative for pushing their users to update the certificate.

Is possible that a mail to each single users will be sent a some point.

As regarding the site, how are you approaching the issue?

Site Report 2024

29 Requests have been sent to Sites/Countries representatives

- 58 sites
- 35 sites providing pledged resources
- 32 GRID Storages
- 5 Tape systems

Thank you very much to all people who worked on it!

NEW CHALLENGES FOR SITES

- Token Based Authentication
- Update the Operative system (RHEL9/Almalinux9)
- Network Operation (Link update, Jumbo Frame)

TYPE	Resource provided	Pledged for 2024 JFY
CPU Pledge	512,6 kHS06/kHS23*	520kHS06
CPU Opport.	412.6 kHS06/kHS23	
DISK	21.292 PB	25 PB
TAPE	12.929 PB	9.520 PB

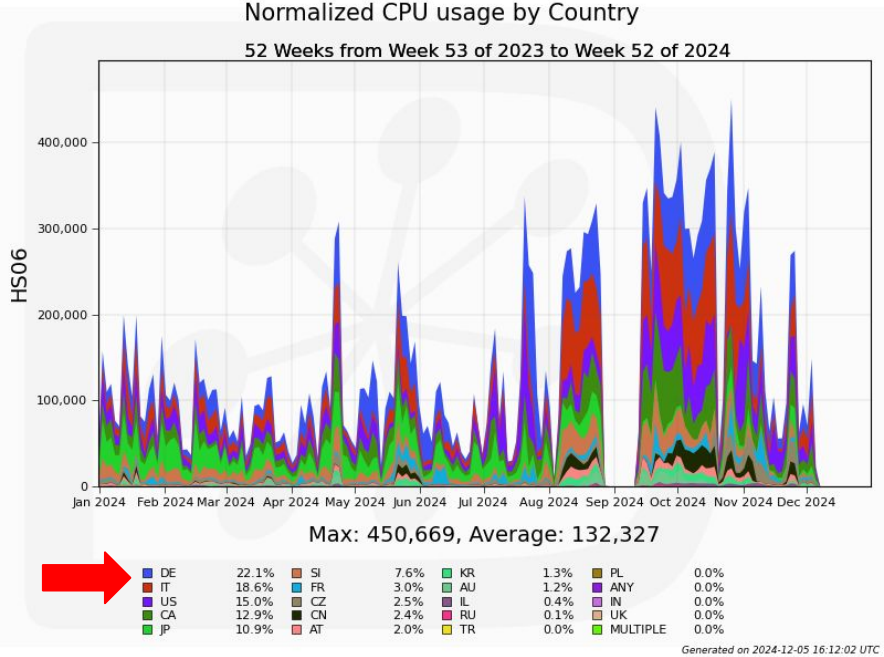
For Production: 34.9 kjobslots pledged + 36 kJobslot opportunistic

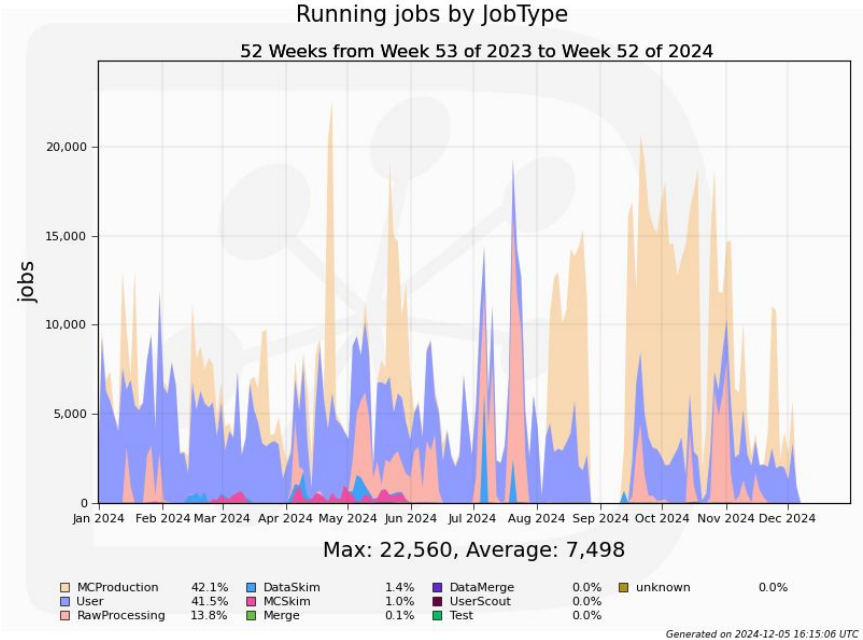
*Including local resources and calibration

TYPE	Resource provided
CPU	36,7 kHS06/HS23
DISK	1.685 PB

Specific Resource for calibration

Activity in 2024





Milestone 2024 - 12%

Italian Share i = 18.6%

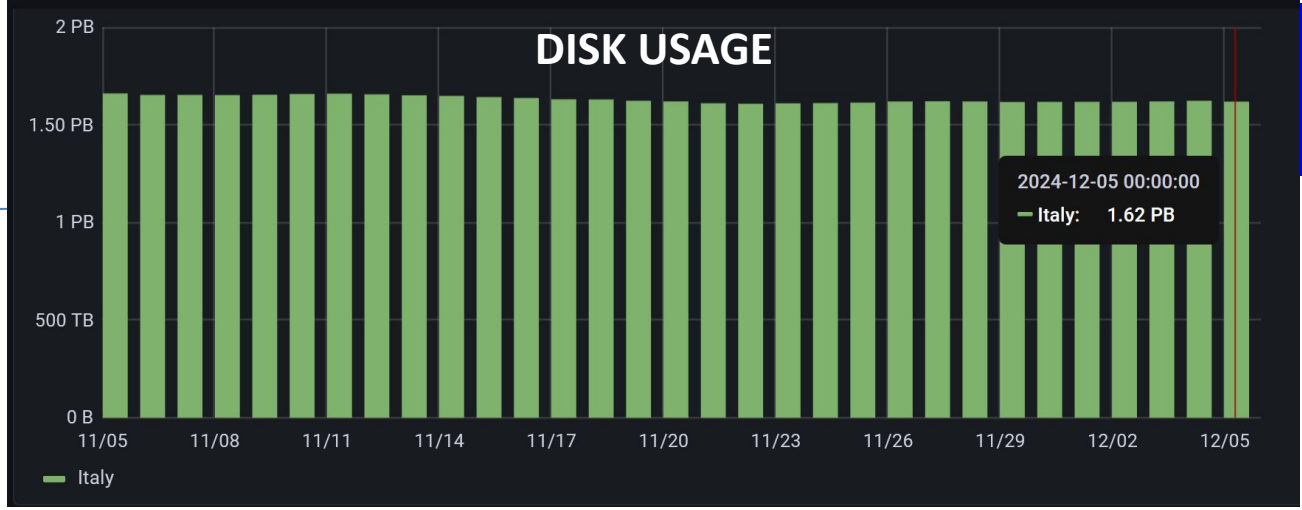
Storage usage

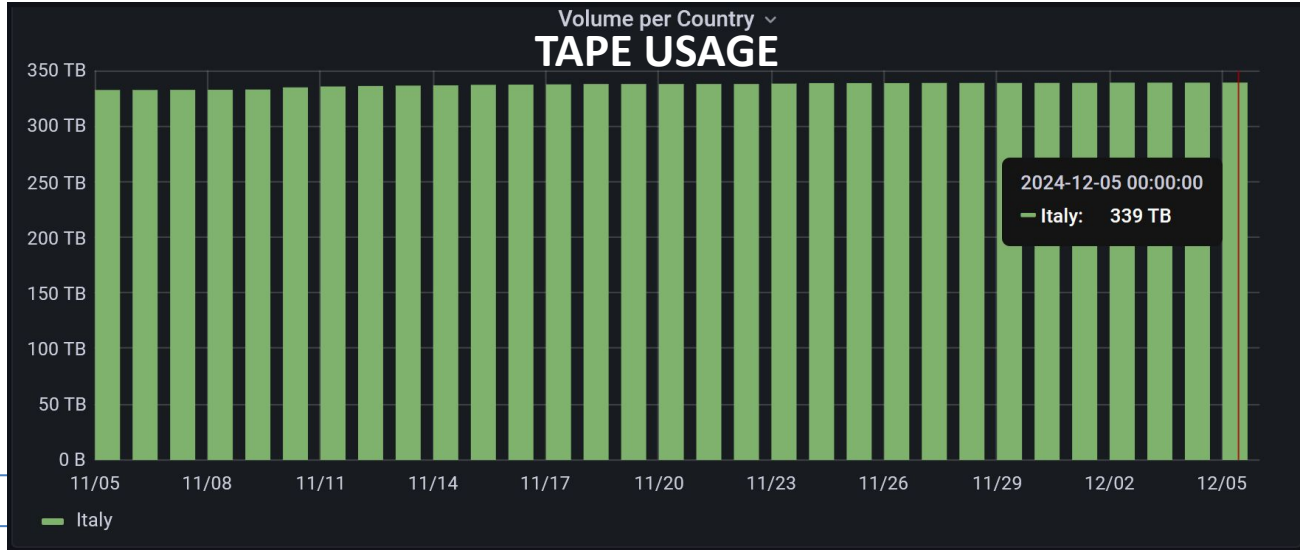
Disk provided

2.230 TB

Tape provided

650 TB





Resoconto risorse pledged per Belle II ad oggi luglio 2024

SITE	CPU	STORAGE	TAPE
CNAF	31kHS06	820TB	650 TB
Napoli+Cosenza	19.9kHS06	860TB	
Pisa	8kHS06	200TB	
Torino	6kHS06	350TB	
TOTALE OGGI	**64.9kHS06**	**2.230TB**	**650TB**

Siamo in attesa che venga fornito il pdedged dello storage al CNAF sia 2023 (300TB) che 2024 (200TB). Previsto a brevissimo

Necessità di calcolo - Giugno 2024

Share italiano MC 2025: 14%

Conservazione 20% della seconda copia dei RAW DATA al CNAF e relativo processing/reprocessing

Resource Estimate: Le risorse di alcuni paesi precedente inclusi tra quelli fornitori di risorse di calcolo, verranno ridistribuite su altre nazioni tra cui l'Italia

	2025	2026	2027	2028
Total tape (PB)	11.8	15	20.4	25.9
Total disk (PB)	19.4	25.3	30.3	37.7
Total CPU (kHS06)	247	492	547	643

	Pledge Italia 2024	Provided	Pledge Italia 2025	Needs
TAPE (TB)	490	650	710	+60
DISCO (TB)	2.522	2.230	2.460	No needs
CPU (kHS06/kHS23)	64,86	64,9	33,17	No needs

Napoli Site

Site of Napoli LCG.Napoli.it

Upgrade to EL9 ongoing.

Two cluster currently running one with CentOS7 and one with Alma9

Around half of INFN resources are migrated to Alma9.

Plan to stay like that for a while.

The universitary cluster will schedule later the update.

HTCondor/Condor Release 23 with SSL authentication.

Site of Napoli LCG.Napoli.it

32 nodes - 96 cores each - INFN Cluster

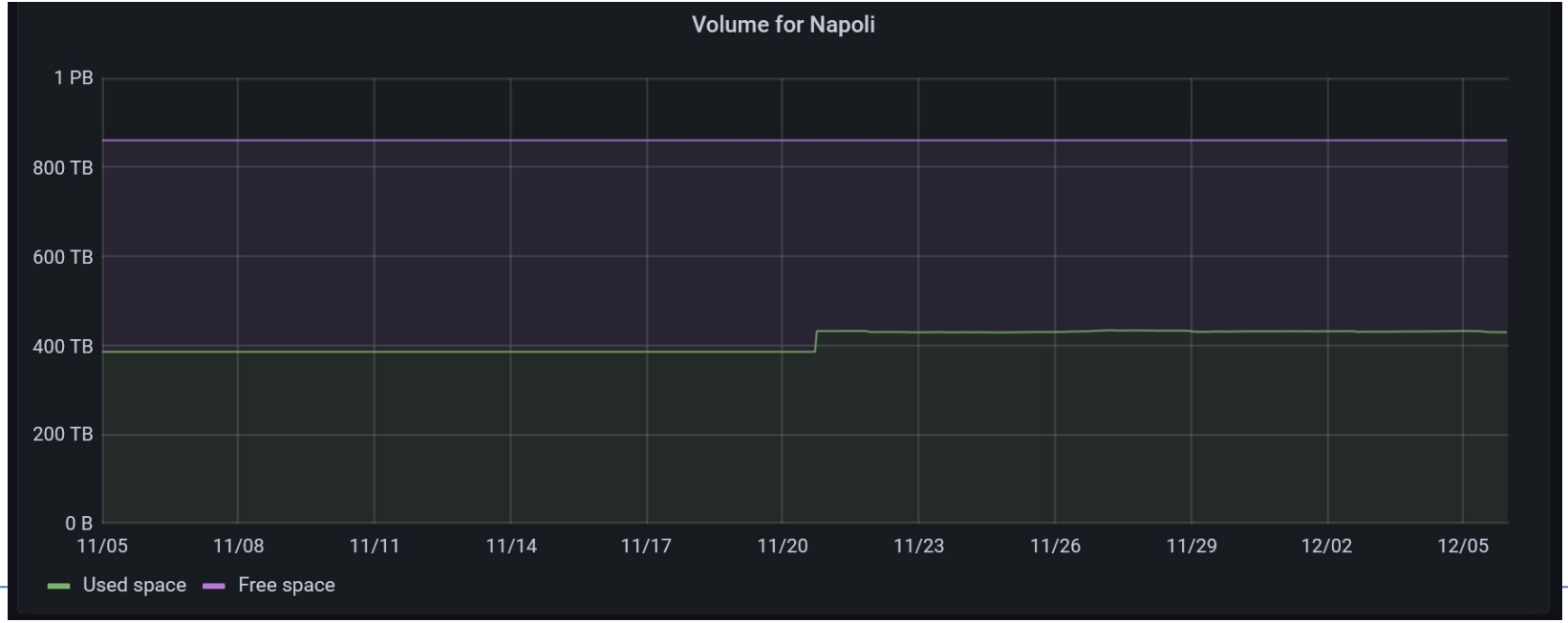
- 20GB disk per core
- 8GB RAM per core
- SET CGROUP_MEMORY_LIMIT_POLICY = none

1000 Cores on the UNINA Cluster

- 20GB disk per core
- 8GB RAM per core

Storage Napoli-TMP-SE/Napoli-DATA-SE

Storage 860 PB available, about half used .



Napoli site: GEANT TCS issue

All certificates will be renewed next week

Plan to issue some certificate with * try to understand how we can use in GRID.

All belle II users have update their certificates.